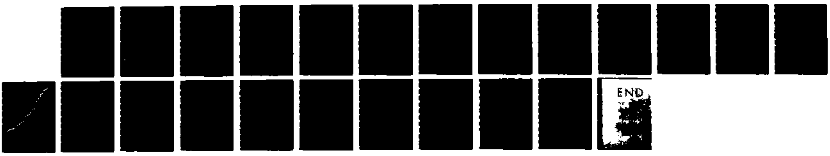
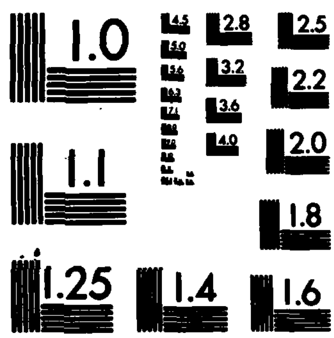
AD-A140 599 INFORMATIVE STOPPING RULES(U) CALIFORNIA UNIV BERKELEY OPERATIONS RESEARCH CENTER R E BARLOW ET AL. JAN 84 ORC-84-1 AFOSR-81-0122 1/1

UNCLASSIFIED F/G 12/1 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6

INFORMATIVE STOPPING RULES

by

Richard E. Barlow† and S. W. W. Shor††

ORC-84-1 January 1984

†Department of Industrial Engineering and Operations Research, University of California, Berkeley, California.

††Bechtel Power Corporation, San Francisco, California.

This research was supported by the Air Force Office of Scientific Research (AFSC), USAF, under Grant AFOSR-81-0122 with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ORC 84-1	2. GOVT ACCESSION NO. AD-A140599	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) INFORMATIVE STOPPING RULES		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Richard E. Barlow and S. W. W. Shor		8. CONTRACT OR GRANT NUMBER(s) AFOSR-81-0122
9. PERFORMING ORGANIZATION NAME AND ADDRESS Operations Research Center University of California Berkeley, California 94720		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS United States Air Force Air Force Office of Scientific Research Bolling Air Force Base, D.C. 20332		12. REPORT DATE January 1984
		13. NUMBER OF PAGES 22
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Statistical Inference
Bayesian Approach
Information
Stopping Rules

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(SEE ABSTRACT)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT

A stopping rule, given data, is informative relative to parameters of interest if it is random and statistically dependent on those parameters. Practical examples considered in detail illuminate the role of informative stopping rules and show how they may arise in practice. The discussion is based on the Bayesian Approach.

INFORMATIVE STOPPING RULES

by

Richard E. Barlow and S. W. W. Shor

Raiffa and Schlaifer (1961) discuss *noninformative* stopping rules in their book. However, the role and importance of *informative* stopping rules, especially relative to censored data, is not made clear. The following examples and discussion clarify the role of informative stopping rules in data analysis.

In recording or extracting information for statistical analysis, some rule or set of instructions must be employed either explicitly or implicitly in order to terminate the recording or information extraction procedure. For example, records on fossil fuel electrical power plants were searched relative to the frequency and duration of forced outages exceeding 60 days. Since the records were tabulated by quarter of a year, all outages exceeding 30 days in a quarter were extracted from the record. If an outage exceeding 30 days was still in effect at the end of a quarter, the following quarter was searched to complete the record for that particular outage. If an outage exceeded 30 days from the start of a quarter, the previous quarter was searched to complete the record also for that particular outage. By following this procedure, we could be sure that no 60 day or greater outage was missed. Relative to 60 day or greater outages, this particular stopping rule, given the data, was noninformative with respect to model parameters. All of our information about model parameters was contained in the number and durations of 60 day or greater outages--none of which were missed.

However, it subsequently became necessary to use the same extracted data to assess the frequency and duration of 30 day or greater forced outages. Relative to these outages, our search procedure, and hence our stopping rule, almost surely missed some 30 day or greater outages in the record. See Figure 1. The missed outages constitute an unobserved nuisance parameter, say ϕ , whose distribution depends on both the stopping rule and the unknown model parameters of interest. Given observed 30 day or greater outages and the knowledge that some could have been missed, our stopping rule was now informative relative to unknown model parameters defining the probability distribution for outage durations exceeding 30 days.

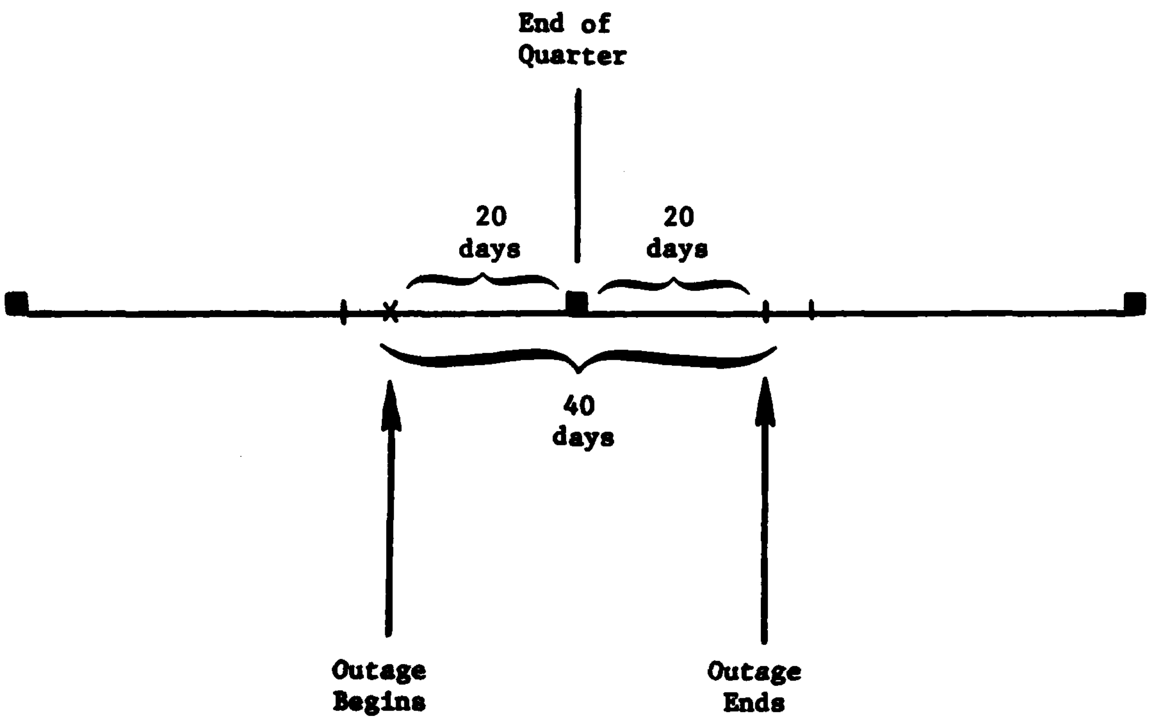


FIGURE 1

EXAMPLE OF A MISSED OUTAGE
EXCEEDING 30 DAYS

1. DEFINITIONS

Suppose a unit lifetime (or downtime) duration, X, depends on unknown parameters, $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_s)$. Observation on a unit may stop before a unit lifetime (or downtime) duration is observed. Let STOP be a *rule* or a *set of instructions* which determines when observation of a unit stops. STOP may be random and dependent on unknown parameters. The stopping rule is *not* necessarily the same as the "stopping time."

The stopping rule discussed in the introduction was: "Extract a downtime duration from a quarterly record if it exceeds 30 days, otherwise ignore it." Consequently, relative to inference about 30 day downtime duration probability parameters, this stopping rule is *random*, since observation of any particular unit downtime is random.

Definition:

A stopping rule, STOP, is noninformative relative to model parameters $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_s)$ if STOP is statistically independent of $\underline{\theta}$, given data; i.e.,

$$\text{STOP} \perp \underline{\theta} \mid \text{Data} .$$

Another way of saying this is that the posterior density for $\underline{\theta}$, given the data, is the same as the posterior density for $\underline{\theta}$ given the data *and* the stopping rule; i.e.,

$$\pi(\underline{\theta} \mid \text{Data}) = \pi(\underline{\theta} \mid \text{Data, STOP})$$

for all $\underline{\theta}$.

If the stopping rule is not random relative to the data, then it is independent of $\underline{\theta}$. In our example, the stopping rule was not random relative to 60 day downtime durations, since none were missed. Consequently, the stopping rule was noninformative in this case.

Formulas for calculating the likelihood have been developed for general sampling plans in which the stopping rule is noninformative given data [cf. Barlow and Proschan (1980), Theorem 1.7]. Most of the stopping rule examples in the statistical literature concern noninformative stopping rules. An exception is a paper by Roberts (1967) which presents an example based on fish capture-recapture sampling methods. However, he comments that "interest in exploiting the information in the stopping rule is likely to be great only for very small sample sizes." Although his statement applies to his examples, it is not true for the example we now discuss in detail.

2. ANALYSIS OF AN INFORMATIVE STOPPING RULE

Consider the stopping rule, STOP , discussed in the introduction and the list in Table 2.1 of outages 30 days or greater extracted from quarterly records. There were $k = 72$ such outages found. However, due to the stopping rule, some outages of this type were almost surely missed.

Since we are interested in the conditional probability distribution of the excess over 30 days of such outages, we subtracted $720 = (24)(30)$ hours from the listed duration hours in Table 2.1. Let $y_1, y_2, \ldots, y_k$ denote these excess downtimes. A transform of this data was then plotted in Figure 2.1. Were the conditional distribution of excess durations exponential, we would expect the plot to lie close to the 45 degree line and, in fact, cross it. Since our plot exhibits this type of behavior and since also the sample coefficient of variation is close to 1, we adopt an exponential model

$$f(x \mid \theta) = (1/\theta)e^{-x/\theta}$$

for our conditional probability distribution of excess durations. The rational for this procedure is based on

(a) the relatively large sample size, $k = 72$;

(b) the total time on test plot is the maximum likelihood estimate of a transform of the sample distribution.

Obviously, the exponential model is mathematically convenient and can be justified if it provides a close approximation measure for our uncertainty about conditional durations, given the data k and $y_1, y_2, \ldots, y_k$.

TABLE 2.1

FOSSIL UNITS 575 MW AND LARGER

DATE	UNIT	DOWNTIME DURATION HOURS
Quarter 1 1976		
2/18/76	Amos Unit 1	1412
1/30/76	H. L. Bowen Unit 1	1018
2/22/76	Kincaid No. 2	1660
2/07/76	Ninemile Point No. 4	1294
Quarter 2 1976		
4/01/76	H. L. Bowen Unit 1	4390
4/01/76	Cardinal Unit 2	792
5/17/76	Monroe No. 1	781
4/20/76	W. H. Sammis No. 6	733
Quarter 3 1976		
7/20/76	Bowline Point Unit 1	1469
8/08/76	Kincaid No. 2	1125
Quarter 4 1976		
10/11/76	H. L. Bowen Unit 2	797
12/20/76	Ninemile Point No. 5	2755
Quarter 1 1977		
3/07/77	Amos Unit 2	2925
1/18/77	Chalk Point Unit 3	806
3/21/77	Cliffside Unit 5	996
2/28/77	Gorgas Unit 10	720
2/05/77	Mohave Unit 2	1161
2/14/77	Ninemile Point No. 4	2548
1/03/77	W. H. Sammis No. 6	4432
Quarter 2 1977		
4/30/77	Astoria Project	3913
4/08/77	Baxter Wilson Unit 2	940
6/24/77	H. L. Bowen Unit 1	1053
5/27/77	Bowline Point Unit 2	1631
4/05/77	Oswego Unit 5	1035
Quarter 3 1977		
8/10/77	Belews Creek Unit 1	773
8/08/77	Chalk Point Unit 3	915
9/30/77	Chalk Point Unit 3	846
7/07/77	Sherburne Unit 1	1521
Quarter 4 1977		
11/06/77	Amos Unit 2	850
11/09/77	Baldwin Unit 2	792
11/30/77	Cumberland Unit 1	766

DATE	UNIT	DOWNTIME DURATION HOURS
11/23/77	Kincaid No. 1	1928
11/15/77	La Cygne Unit 1	961
10/04/77	W. H. Sammis No. 7	1257
Quarter 1 1978		
2/24/78	Cumberland Unit 1	851
1/30/78	Harrison Unit 2	3625
2/01/78	Mohave Unit 1	3528
3/10/78	Ninemile Point No. 5	2216
1/09/78	W. H. Sammis No. 7	3109
Quarter 2 1978		
5/05/78	Gaston Steam Plant Unit 5	864
5/15/78	Marshall No. 3	1408
5/19/78	Ninemile Point No. 4	2958
5/03/78	Tradinghouse Creek Unit 2	2188
Quarter 3 1978		
7/01/78	Centralia Unit 1	1559
7/16/78	Ninemile Point No. 4	3557
9/29/78	Ormond Beach Unit 2	776
Quarter 4 1978		
11/06/78	Keystone No. 1	768
10/03/78	Oswego Unit 5	1247
Quarter 1 1979		
2/17/79	Conesville Unit 4	2411
1/01/79	Hatfield No. 1	4167
1/01/79	Hatfield No. 1	3320
1/01/79	Mt. Storm No. 1	2159
2/03/79	Paradise No. 1	3424
3/30/79	Ravenswood No. 3	2903
Quarter 2 1979		
4/01/79	Baxter Wilson Unit 2	1672
6/28/79	Harrison Unit 2	3858
4/01/79	La Cygne Unit 2	1287
5/27/79	Mohave Unit 1	792
Quarter 3 1979		
7/26/79	Astoria Project	1034
8/06/79	Harrison Unit 1	5116
8/30/79	Hudson No. 2	761
8/20/79	Keystone No. 1	1003
7/08/79	La Cygne Unit 1	920
7/23/79	Ormond Beach Unit 2	799
9/22/79	Pittsburg Unit 7	1573
9/29/79	W. F. Wyman Unit 4	1815

DATE	UNIT	DOWNTIME DURATION HOURS
Quarter 4 1979		
10/01/79	Centralia Unit 2	1278
11/22/79	Mt. Storm Unit 3	2586
Quarter 1 1980		
1/12/80	Hatfield No. 1	1287
Quarter 2 1980		
5/05/80	Belews Creek Unit 2	1360
5/14/80	Kincaid No. 1	903
5/04/80	W. H. Sammis No. 7	1286

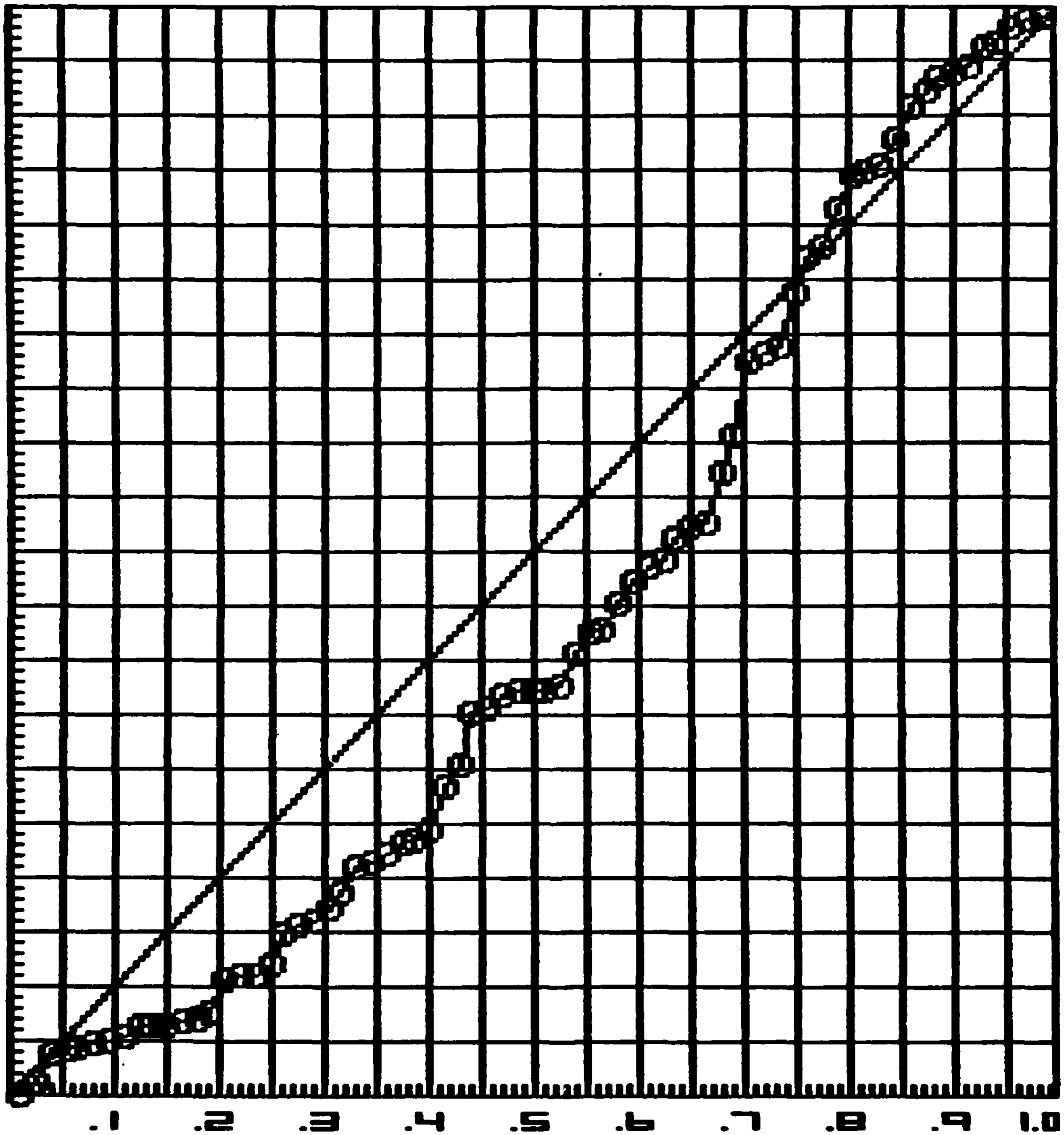


FIGURE 2.1

TOTAL TIME ON TEST PLOT OF OUTAGE DURATIONS IN EXCESS OF 30 DAYS

N	SAMPLE MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION
72	1048.93 hours	1124.5	1.07

Relative to our data base, outages greater than 30 days are fairly rare. Hence, given $t = 492.5$ unit years operating experience, our a priori probability for observing k such outages is

$$P[N(t) = k \mid t,\lambda] = (\lambda t)^k e^{-\lambda t}/k! ,$$

where λ is the expected number of such outages per unit year. We suppose that any exchangeable collection of additional units will have this same unknown rate λ .

3. LIKELIHOOD DERIVATION

Given t, STOP, and the data in Table 2.1, we need to calculate the likelihood for θ and λ. Let ℓ be the calendar period from January 1, 1976 to the end of the second quarter 1980 less 30 days, since we would not have caught such outages beginning within 30 days of the end of the second quarter 1980. Our a priori expectation for the number of such outages occurring in ℓ is λt, given λ. The conditional probability that such an outage, having occurred in ℓ, will not be missed is

$$\left[1 - \frac{m\Delta}{\ell} p(\theta)\right]$$

where $\frac{m\Delta}{\ell}$ is the probability that such an outage occurs in a $\Delta = 30$ day interval preceding the end of a quarter and $m = 17$ is the number of critical intervals. This probability is multiplied by $p(\theta)$, the conditional probability that such an outage *falling within a critical interval* will actually be missed. Hence, our prior expectation for the number of observed outages in t unit years operating experience is

$$\lambda t\left[1 - \frac{m\Delta}{\ell} p(\theta)\right] .$$

We now derive the formula for $p(\theta)$. Suppose an outage of length Z starts in a critical $\Delta = 30$ day interval and at time x, u time units from the end of a quarter. See Figure 3.1. Let $Z = \Delta + Y$ where Y is the excess over 30 days. Then

$$p(\theta) = \int_0^\Delta P\{Z \leq u + \Delta \mid Z > \Delta \text{ , outage starts in } \Delta \text{ interval}\} \frac{du}{\Delta}$$

$$= \int_0^\Delta P\{Y \leq u\} \frac{du}{\Delta} = \int_0^\Delta (1 - e^{-u/\theta}) \frac{du}{\Delta}$$

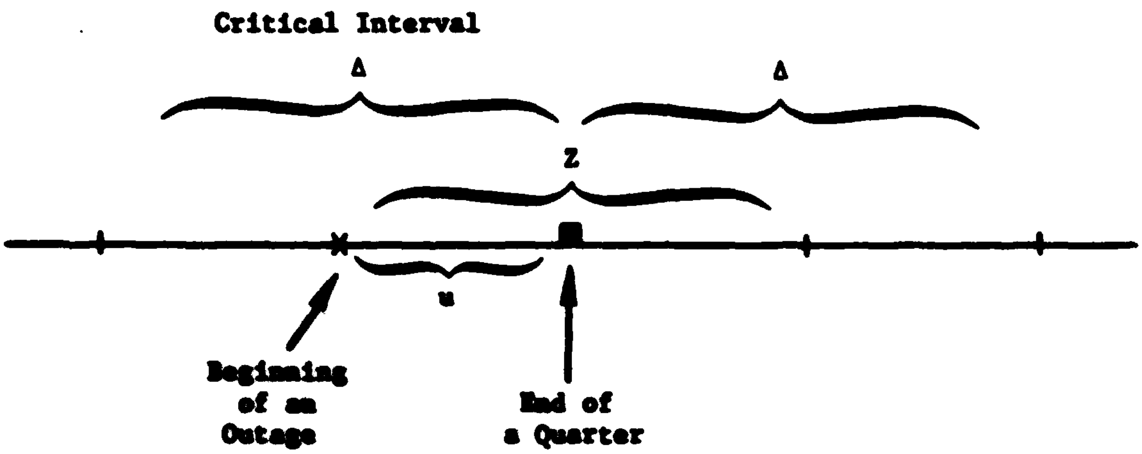


FIGURE 3.1

DIAGRAM ILLUSTRATING A MISSING OUTAGE

and

$$p(\theta) = 1 - \frac{\theta}{\Delta}\,(1 - e^{-\Delta/\theta}) \ .$$

The observed number of outages is also Poisson with parameter $\lambda t\left[1 - \frac{m\Delta}{\ell}\,p(\theta)\right]$. This can be shown from first principles by successively conditioning and unconditioning starting with the formula

$$P[N(t) = k \mid \lambda,\theta,t,\mathrm{STOP}]$$

$$= \sum_{n=k}^{\infty} \left\{ \sum_{j=n-k}^{n} \binom{n}{j}\left(\frac{m\Delta}{\ell}\right)^{j}\left(1 - \frac{m\Delta}{\ell}\right)^{n-j}\binom{j}{n-k}[p(\theta)]^{n-k}[1-p(\theta)]^{j-n+k} \right\} \frac{(\lambda t)^{n} e^{-\lambda t}}{n!} \ .$$

By straightforward algebra, we obtain

$$P[N(t) = k \mid \lambda,\theta,t,\mathrm{STOP}]$$

$$= \frac{(\lambda t)^{k}\left[1 - \frac{m\Delta}{\ell}\,p(\theta)\right]^{k} e^{-\lambda t\left[1 - \frac{m\Delta}{\ell}\,p(\theta)\right]}}{k!} \ .$$

The likelihood for λ and θ, given k observed such outages with excess durations $y_1, y_2, \ldots, y_k$ and stopping rule STOP, is

$$L(\lambda,\theta \mid k, y_1, \ldots, y_k, t, \mathrm{STOP}]$$

$$\prec \lambda^{k}\left[1 - \frac{m\Delta}{\ell}\,p(\theta)\right]^{k} e^{-\lambda t\left[1 - \frac{m\Delta}{\ell}\,p(\theta)\right]} \theta^{-k} e^{-T/\theta}$$

since conditional on k outages in ℓ,

$$L(\theta \mid k,T) \prec \theta^{-k} e^{-T/\theta}$$

where $T = \sum_{i=1}^{k} y_i$. Note that conditional on k, (k,T) is sufficient for θ under the exponential model.

Were the stopping rule STOP given data noninformative, the likelihood would have been

$$\lambda^k e^{-\lambda t} \theta^{-k} e^{-T/\theta} .$$

In this case, we would have estimated λ and θ by $\lambda^* = k/t$ and $\theta^* = T/k$. However, a closer approximation to the MLE's can be found by using $\hat{\theta} = T/k$ and calculating the value λ for which

$$\lambda^k \left[1 - \frac{m\Delta}{\ell} p(\hat{\theta})\right]^k e^{-\lambda t \left[1 - \frac{m\Delta}{\ell} p(\hat{\theta})\right]}$$

is maximum. This approximate MLE, $\hat{\lambda}$, is

$$\hat{\lambda} = \frac{k}{t\left[1 - \frac{m\Delta}{\ell} p(\hat{\theta})\right]} .$$

[See DeGroot (1970, p. 199) for Bayesian justification for MLE.]

<u>Numerical Example</u>:

For the data in Table 2.1, $k = 72$, $T = 7.5768 \times 10^4$ hours and $t = 492.5$ unit years so that $\lambda^* = 0.146$ per unit year. On the other hand, $\hat{\theta} = T/k$ and $p(\hat{\theta}) \doteq 0.276$ so that

$$\hat{\lambda} = 72/492.5\left[1 - \frac{17(30)}{\ell} (0.276)\right]$$

and

$$\hat{\lambda} \doteq 72/(492.5)(0.911) \doteq 0.163 .$$

This is a 9% increase over λ^* !

4. STOPPING RULES USED IN LIFE TABLE ANALYSIS

Breslow and Crowley (1974) and Lindley (1979) studied the following model relative to estimating death rates for specified age intervals. Associated with each individual is a pair of independent random quantities, X , the lifetime and Y , the withdrawal time. The raw observations for an individual are $Z = \text{Min}\,(X,Y)$, the time at which he leaves either through death or withdrawal, and an indicator which says whether the departure was caused by death or withdrawal. The time scale is then divided into nonoverlapping intervals and the Z's grouped so that observation is only made on the interval within which he left the system. The quantities for different individuals are judged independent and identically distributed. Consequently, if N individuals are present at the beginning of an interval, the data consists of D , the number who were observed to die in the interval; W the number observed to withdraw alive during the interval; and S the number who survived to enter the next interval. We consider only the single interval $[0,\Delta]$. Let X have distribution F and let $\phi = F(\Delta)$ be the random quantity of interest. Let Y have distribution H , $\theta = H(\Delta)$ and

$$\rho = \int_0^\Delta F(x)dH(x)/\theta\phi$$

so that ρ is the conditional probability that a death is observed, given that both withdrawal and death take place in the interval. As Lindley (1979) points out, the likelihood for ϕ , θ and ρ can be calculated given D , S and W . Obviously, θ and ρ are nuisance parameters.

Since observation on a unit ceases at $\text{Min}\,(X,Y,\Delta)$ and Y is random, the stopping rule is random. Is the stopping rule informative?

Given the stopping rule, W provides partial information about ϕ. Note that $W = D^* + S^*$ where D^* is the number of withdrawals that would have been observed to die in $[0,\Delta]$ had they not been withdrawn and similarly for S^*. The probability that an individual will die in $[0,\Delta]$ and his death will not be observed is $\theta\phi[1 - \rho]$, which clearly depends on ϕ. The likelihood as calculated by Lindley (1979) is

$$L(\phi,\theta,p \mid D,S,W)$$
$$= \phi^D(1 - \rho\phi)^W(1 - \phi)^S\{1 - (1 - \rho)\theta\}^D\theta^W(1 - \theta)^S .$$

Since θ and ρ are nuisance parameters, they must be integrated out with respect to a joint prior for (ϕ,θ,ρ).

Were we given the ages of death $(x_1, \ldots, x_k)$, the survival and withdrawal ages $(\ell_1, \ldots, \ell_m)$ and were the parameter of interest the force of mortality $[r(u),u \geq 0]$, then the likelihood would be

$$L(r(u),u \geq 0 \mid x_1, \ldots, x_k , \ell_1,\ell_2, \ldots, \ell_m)$$
$$\prec \left[\prod_{i=1}^{k} r(x_i)\right] e^{-\int_0^\infty n(u)r(u)du}$$

where $n(u)$ is the number of individuals observed surviving to age u. In this case, the stopping rule would be noninformative, since the posterior for $[r(u),u \geq 0]$ would not depend on the stopping rule.

5. CONCLUSION

In analyzing data, it is important to think about the way in which the data was obtained. If the stopping rule, given data, is informative in the sense defined in Section 1, the likelihood calculation may be more difficult, but resulting estimates may differ significantly from those which ignore the stopping rule. Whether or not the information contained in the stopping rule is relevant depends on the observed data as well as the model, the parameter of interest and its prior probability as the previous examples demonstrate.

REFERENCES

Barlow, R. E. and F. Proschan (1979), "Inference for the Exponential Life Distribution," ORC 79-16, Operations Research Center, University of California, Berkeley.

Barlow, R. E. and F. Proschan (1980), "Life Distribution Models and Incomplete Data," ORC 80-20, Operations Research Center, University of California, Berkeley.

Barlow, R. E. and F. Proschan (1981), STATISTICAL THEORY OF RELIABILITY AND LIFE TESTING, To Begin With, Silver Springs, Maryland.

DeGroot, M. (1970), OPTIMAL STATISTICAL DECISION, McGraw-Hill, New York.

Lindley, D. V. (1979), "Analysis of Life Tables with Grouping and Withdrawals," Biometrics, Vol. 35, pp. 605-612.

Raiffa, H. and R. Schlaifer (1961), APPLIED STATISTICAL DECISION THEORY, The M.I.T. Press, Cambridge, Massachusetts.

Roberts, H. V. (1967), "Informative Stopping Rules and Inference About Population Size," Journal of the American Statistical Association, Vol. 62, pp. 763-775.

FILME